MARYVILLE UNIVERSITY

Predictive Modeling

ACSC412/512

Fall 2017

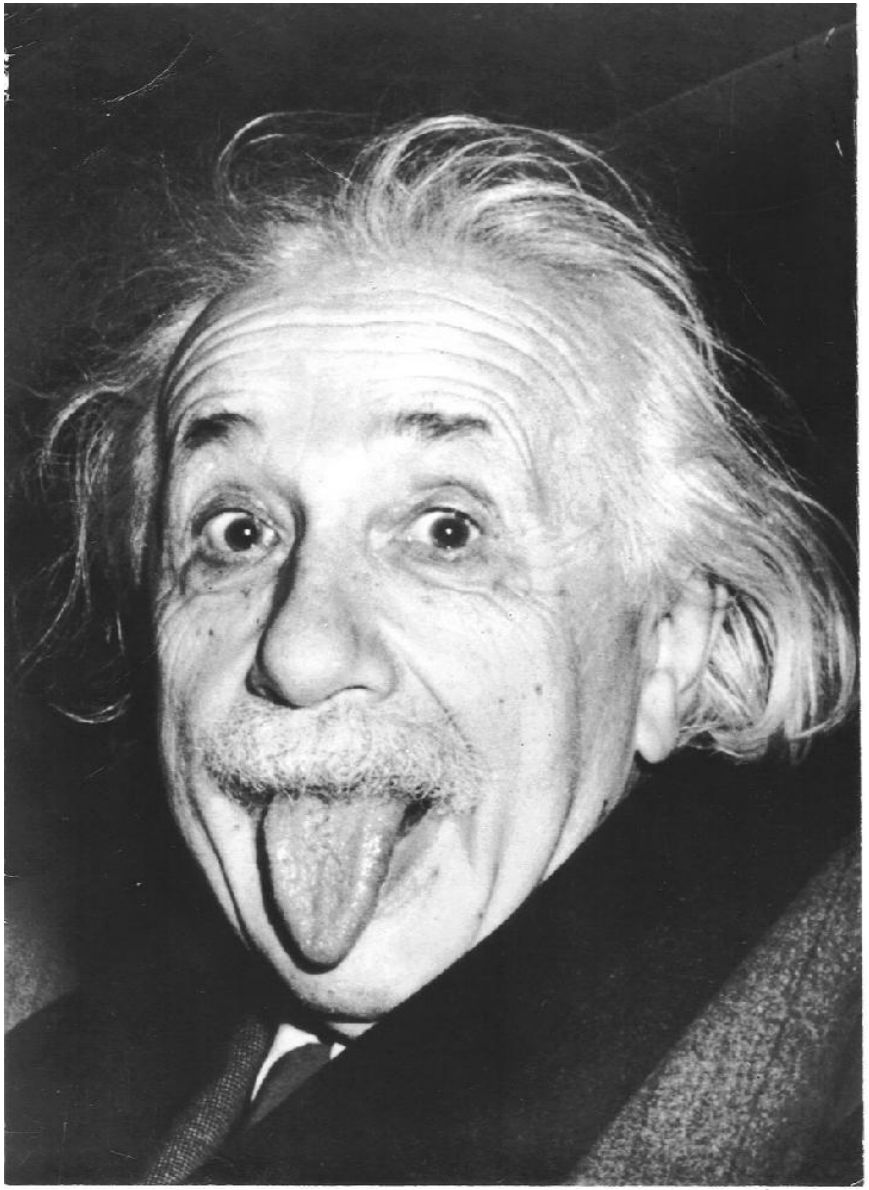
Richard Xu

- About me – Richard (Qichun) Xu, PhD, FSA
  - education background
  - work experience
- Course syllabus
  - Goals & Over view
  - Time, schedule
  - Homework; communication
  - Question?
- Today's contents
  1. Predictive Modeling Introduction
  2. Review of Ordinary Least Squares & maximum likelihood; why OLS is not effective in insurance
  3. Introduction of R

Introduction

Know where to find the information and how to use it - That's the secret of success

Albert Einstein



What is Predictive Modeling?

What?

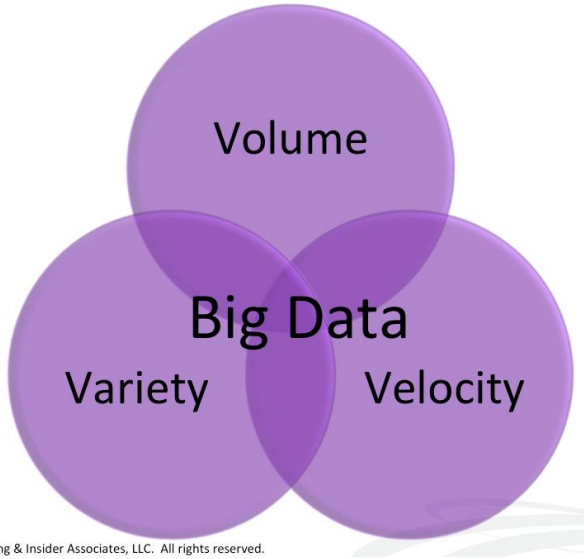
1. **Data**
   High quality data
2. **Modeling**
   Statistical model
3. **Prediction**
   Business decisions

Why?

- Grow Business
  - Identify potential customers, new & existing
  - Create demand of insurance products
  - Enhance sales process
- Improve efficiency
  - Accurate & granular view of driven risk factors
  - Better risk selection
  - Improve pricing capabilities
- Create an advantage hard for competitors to replicate

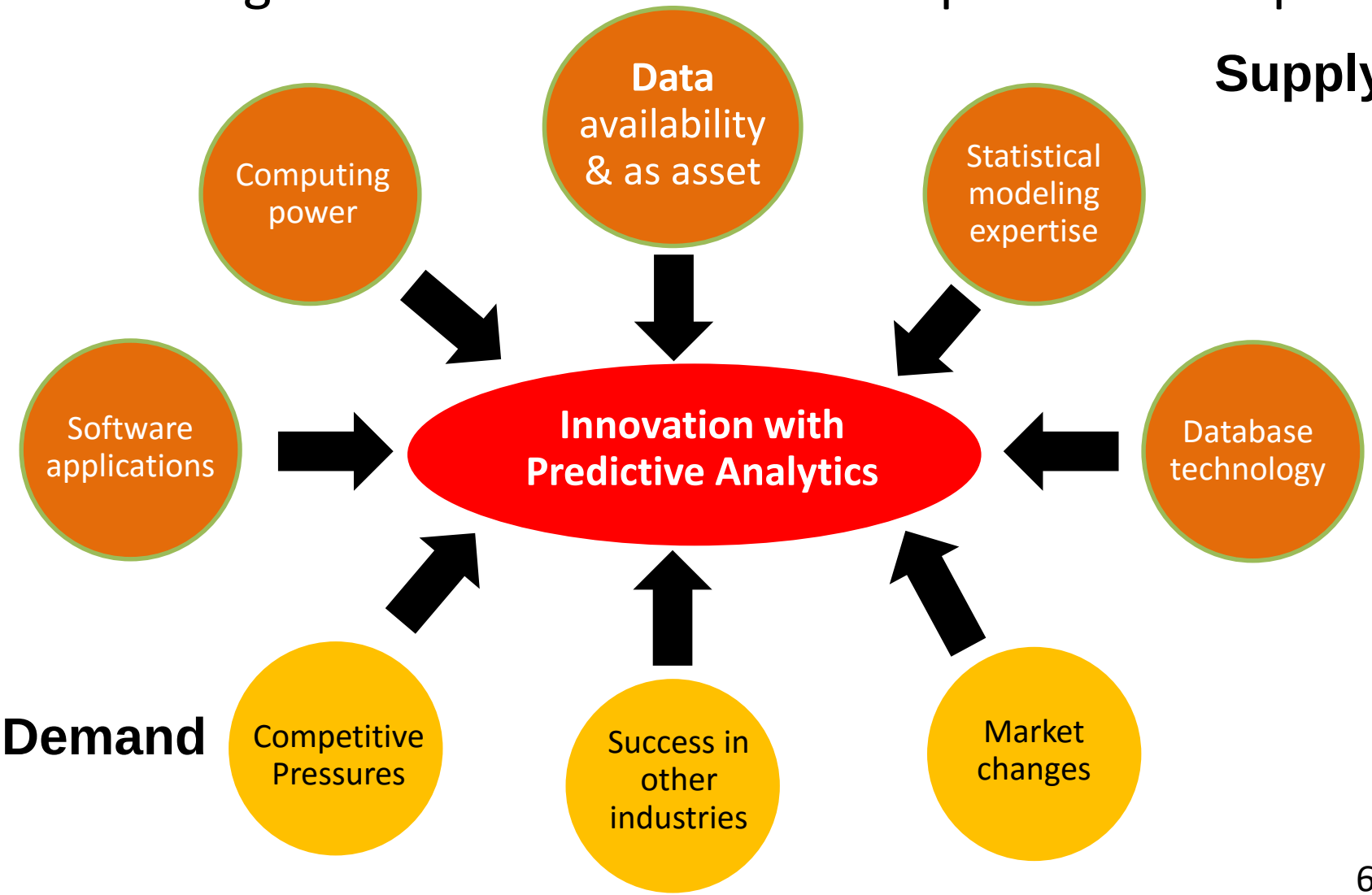
PM is about statistics, but more about data & business

- Predictive Modeling – a process by which a statistical model is created on historical data to best predict the probability of a future outcome
- What about "data mining"?
- Heard about "Big Data"?
  - 3Vs – Volume, Velocity, Varity
  - Structured vs. unstructured
  - How about insurance application?
- Why statistical modeling?
- Actuary, data scientist, etc.
- Correlation vs. causation



Why are Predictive Analytics becoming so popular?

Increasing ease of access to enablers plus demand pull



Data Sources

MARYVILLE UNIVERSITY

Financial & Socioeconomic Data

- Property Ownership
- Income & Wealth
- Banking Data
- Financial Credit

Medical Data

- Prescription Histories
- Laboratory Profiles
- Third-Party Data (MIB)
- Physician Reports/EHR

(Potential) Customer

Insurance Data

- Application Disclosures
- Underwriting Data
- Policy values
- Claim/Lapse Experience

- Consumer Spending
- Driving Violations
- Hobbies & Interests
- Physical Activity

Behavioral & Lifestyle Data

Build Models & Make Predictions

- ✓ Likelihood to buy new product
- ✓ Probability to qualify for new product by SI/GI
- ✓ Likelihood to lapse
- ✓ Likelihood to be fraudulent
- ✓ …and more

Opportunity

Underwriting

Identify best risks

Be fast & consistent

Prioritize cases

Reduce not-taken rates

Claims

Predict claim frequency

Identify claim severity

Prioritize resources

Identify claims most likely fraudulent/rescinded

Pricing/ Reserves

Improve pricing accuracy

Identify deviation of pricing variables

Reserve more accurate

Compute reserve variance

Experience Analysis

Identify drivers in experience

Handle low credibility data

Create own mortality/lapse tables

Sales & Marketing

Make effective campaigns

Recommend products

Select new agents

Monitor existing agents

In Force Business

Predict lapses

Design retention strategies

Offer other products

- Regression study
  - Linear(OLS), general linear
  - **Generalized LM (GLM)**
  - Survival Models (Cox Proportional Hazard)
  - Generalized Additive Models (GAM)
  - Multilevel/Hierarchical Linear Model(HLM)
- Time series analysis
- Some other advanced tools
  - Data clustering – similarity between data
  - Decision tree based: CART, Random Forest, MARS, Ada-Boosting, etc.
  - Other machine learning algorithms
    - Neural network, Genetic programming, Support vector machine, Bayesian inference, Cluster analysis, K-nearest neighbor

Terminology

- Supervised vs. Unsupervised Learning
  - Supervised: estimate expected value of *Y* given values of *X*.
    - GLM(OLS), Cox, CART, MARS, Random Forests, SVM, NN, etc.
  - Unsupervised: find interesting patterns amongst X; no target variable
    - Clustering, Correlation / Principal Components / Factor Analysis
- Classification vs. Regression
  - Classification: to segment observations into 2 or more categories
    - GLM, CART, RF, SVM, NN
  - Regression: to predict a continuous amount.
    - GLM(OLS), COX, SVM,
- Parametric vs. Non-Parametric
  - Parametric Statistics: probabilistic model of data
    - GLM
  - Non-Parametric Statistics: no probability model specified
    - classification trees, NN